Optimization of polling systems with Bernoulli schedules

Blanc, J.P.C.; van der Mei, R.D.

Published in:
Performance Evaluation

Publication date:
1995

ELSEVIER

Performance Evaluation 22 (1995) 139–158

PERFORMANCE EVALUATION
An International Journal

Optimization of polling systems with Bernoulli schedules

J.P.C. Blanc *, R.D. van der Mei

Faculty of Economics, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands

Received June 1992; revised September 1993

Abstract

Many computer-communication networks in which the transmission right is circulated among the nodes have been modeled as polling systems. This paper concerns optimization of cyclic polling systems with respect to the service disciplines at the nodes. The service disciplines are chosen to be Bernoulli schedules. Because the optimization problem is not analytically tractable, a numerical approach to determine the optimal schedule, based on the power-series algorithm, is discussed. Light- and heavy-traffic asymptotes of the optimal schedule are presented; they are based on light-traffic asymptotes of the mean waiting times and the stability condition, respectively. A partial solution of the optimization problem is given; this follows directly from the μc-rule for priority systems. The influence of system parameters on the optimal Bernoulli schedule is examined. Finally, a fast approach to approximate the optimal schedule is presented and tested.

Keywords: Polling systems; Optimization; Bernoulli schedules; Power-series algorithm

1. Introduction

A polling system is a multi-queue model, attended to by a single server. Polling models arise naturally in the modeling of many computer-communication networks, where several users compete for access to a common resource, e.g., a central computer or a transmission channel. The reader is referred to [21] and [29] for overviews of the large variety of applications. Because the server usually has no global information on the queue lengths, often a cyclic polling strategy is chosen for such networks, and, moreover, the server continues to move along the queues when the network is empty. In this paper optimization of continuous-time cyclic polling systems with a Bernoulli service strategy at all queues will be considered. The Bernoulli service discipline was introduced for a GI/G/1 vacation model in [18] and applied to cyclic polling systems in [28]. A Bernoulli service strategy, with parameter q_i $(0 \leq q_i \leq 1)$ for queue i, works as follows. When the server arrives at a queue, at least one customer is served, if present;

* Corresponding author.

SSDI 0166-5316(93)E0045-7

otherwise, the server directly proceeds to the next queue. After completion of a service at queue i, the server starts to serve another customer at queue i with probability q_i, if queue i has not yet been emptied; otherwise, the server proceeds to the next queue. Note that in the case $q_i = 0$, queue i is served according to the 1-limited service discipline; if $q_i = 1$, queue i is served exhaustively. The optimization problem considered in this paper is to find a combination of the Bernoulli parameters which minimizes a weighted sum of the steady-state mean waiting times at the various queues, with arbitrary strictly positive weights for each queue. In general, the optimization problem is not analytically tractable. Therefore, we shall discuss some properties of the optimal Bernoulli schedule and moreover, propose a numerical approach to compute the optimal schedule accurately, based on the use of the power-series algorithm (PSA) (cf. [17,1–5]). As this approach may be rather time and memory consuming, we finally present an approximation method for finding the optimal Bernoulli schedule, for which the time and memory requirements are negligible and which yields fairly accurate results over a wide range of admissible parameter values.

The Bernoulli service discipline stochastically limits the number of customers served during one visit of the server to a particular queue through a series of random decisions; here, the Bernoulli parameters may serve as control parameters. The Markovian character of the Bernoulli strategies simplifies the analysis of the model. Getting an insight into such a system may serve as a set-up for the analysis of models with more complex (non-Markovian) service strategies, such as the limited service strategies; under a limited service strategy, the number of customers served during one visit of the server to a particular queue has a fixed upper bound. A main advantage of the Bernoulli strategy is the fact that the range of feasible values of the control parameters is a compact set (in contrast to the limited service discipline, of which the parameters are integers). However, a disadvantage of Bernoulli schedules may be their stochastic character. The reader is referred to [2–6,18,25,28,30,31] for references on models with Bernoulli schedules.

Fuhrmann [13] classifies service strategies into two classes, depending on whether or not the service strategies satisfy a special property, and he shows that models in which all service strategies satisfy this property are relatively easy to analyze. For such models, Resing [27] gives exact expressions for the generating function of the joint queue length at polling instants. However, detailed exact analysis for polling systems with service disciplines which do not satisfy this property is restricted to two-queue models. In general, the Bernoulli service discipline does *not* satisfy this property. As a consequence, there is a need for numerical techniques to analyze queueing systems with Bernoulli schedules. The PSA is a tool for the numerical analysis of queueing models in which the joint queue length process has the structure of a multi-dimensional quasi birth-death process. The PSA is based on power-series expansions of the state probabilities and the moments of the joint queue length distribution as function of the load of the system in light-traffic (cf. [17,1–5]). The reader is referred to [4] for a validation of the PSA by simulation. Leung [20] proposes another numerical technique based on discrete Fourier Transforms to solve the waiting time distributions in polling models with a so-called probabilistically-limited service strategy. The main disadvantage of both algorithms is that the time and memory requirements increase exponentially with the number of queues, so that their use is restricted to fairly small systems. Although the wide variety in routing mechanisms and service strategies opens possibilities for efficient operation, optimization of polling systems has re-

ceived relatively little attention in the literature. The reader is referred to [7,9,10,16,22–24,32] for results on optimization of polling systems.

The rest of the paper is organized as follows. In Section 2 a detailed description of the model will be given and the optimization problem will be formulated. In Section 3 the PSA will be used to derive algebraic expressions for the light-traffic asymptotes of the individual mean waiting times, which yield light-traffic asymptotes of the optimal Bernoulli schedule. For systems with non-negligible switch-over times, the stability condition induces a lower bound on the range of possible values of the components of the optimal Bernoulli parameter. Because for increasing load of the system this lower bound increases to one for each queue, it follows that the optimal service strategy tends to the exhaustive strategy for all queues when the offered load tends to the boundary of the stability region. Moreover, a partial solution of the optimization problem follows directly from the well-known μc-rule for priority systems (cf. [24]). In Section 4, a numerical approach to compute the optimal Bernoulli schedule accurately will be discussed; the approach combines the use of the PSA with the conjugate gradient method. Subsequently, the influence of system parameters on the optimal Bernoulli schedule will be discussed in Section 5. As the numerical approach discussed in Section 4 may be time and memory consuming, we shall propose and test in Section 6 an approximation method for finding the optimal Bernoulli schedule, for which the time and memory requirements are negligible.

2. Model description and problem formulation

Consider a queueing system consisting of s queues, attended to by a single server. Customers arrive at queue i according to a Poisson process with rate λ_i, $i = 1,\ldots,s$. Each queue may contain an unbounded number of customers. The queues are visited in a cyclic order. The number of customers which is served during a visit to a certain queue is determined by a Bernoulli schedule $\boldsymbol{q} = (q_1,\ldots,q_s)$, where q_i is the Bernoulli parameter associated with queue i, $i = 1,\ldots,s$. At each queue the customers are served on a first-in-first-out basis. The service times are assumed to be i.i.d. random variables with finite first and second moments. Let $\beta^{(k)} = (\beta_1^{(k)},\ldots,\beta_s^{(k)})$, where $\beta_i^{(k)}$ denotes the kth moment of the service times at queue i, $i = 1,\ldots,s$, $k = 1, 2$. The times needed by the server to switch from one queue to the next are also assumed to be i.i.d. random variables with finite first two moments. Let $\sigma^{(k)} = (\sigma_1^{(k)},\ldots,\sigma_s^{(k)})$, where $\sigma_i^{(k)}$ denotes the kth moment of the switch-over times from queue $i-1$ to queue i, $i = 1,\ldots,s$, $k = 1, 2$ (with the convention that $\sigma_1^{(k)}$ corresponds to switch-over times from queue s to queue 1). The switch-over times are also allowed to be equal to 0 a.s. here. If the system is empty, the server continues to move from queue to queue. All service times, switch-over times and interarrival times are assumed to be mutually independent. The sum of the arrival processes at the various queues is a Poisson process with rate

$$\Lambda := \sum_{i=1}^{s} \lambda_i. \tag{1}$$

The first two moments β_1 and β_2 of the service time distribution of an arbitrary customer are given by

$$\beta_1 := \frac{1}{\Lambda}\sum_{i=1}^{s}\lambda_i\beta_i^{(1)}, \qquad \beta_2 := \frac{1}{\Lambda}\sum_{i=1}^{s}\lambda_i\beta_i^{(2)}. \tag{2}$$

The *offered load* to queue i, ρ_i, and the *total offered load* to the system, ρ, are defined by

$$\rho_i := \lambda_i\beta_i^{(1)}, \quad i = 1,\ldots,s, \qquad \rho := \sum_{i=1}^{s}\rho_i. \tag{3}$$

Because ρ will be used as a variable in the PSA, define the following quantities:

$$\eta_i := \rho_i/\rho, \quad a_i := \lambda_i/\rho, \qquad i = 1,\ldots,s, \tag{4}$$

referred to as the *relative load* and the *relative arrival rate* of queue i, $i = 1,\ldots,s$; let $a := (a_1,\ldots,a_s)$. The first two moments of the total switch-over time during one cycle of the server along the queues are given by

$$\sigma_1 := \sum_{i=1}^{s}\sigma_i^{(1)}, \qquad \sigma_2 := \sum_{i=1}^{s}\sigma_i^{(2)} + 2\sum_{i=1}^{s}\sum_{j=1}^{i-1}\sigma_i^{(1)}\sigma_j^{(1)}. \tag{5}$$

Necessary and sufficient conditions for the stability of the system have been heuristically derived in [19] and rigorously proved in [12]. For the present model with Bernoulli schedules these conditions read:

$$\chi := \rho + \sigma_1\{\lambda_i(1-q_i)\} < 1, \quad i = 1,\ldots,s. \tag{6}$$

In the sequel it is assumed that the stability condition (6) is satisfied and that the system is in steady-state. The reader is referred to [2–5] for a description of the PSA for this model.

The optimization problem is to minimize the cost function, defined by

$$C(\boldsymbol{q}) := \sum_{i=1}^{s} c_i EW_i, \tag{7}$$

over all $\boldsymbol{q}$ for which (6) holds; here, the components of $c := (c_1,\ldots,c_s)$ are arbitrary strictly positive cost coefficients and EW_i, $i = 1,\ldots,s$, are the steady-state mean waiting times at the various queues, which depend on $\boldsymbol{q}$. The optimal Bernoulli schedule will be denoted by $\boldsymbol{q}^*$. Without loss of generality, it is assumed that the cost coefficients are normalized such that

$$\sum_{i=1}^{s} c_i = 1. \tag{8}$$

In general, the optimization problem is not analytically tractable. Therefore, in Section 3 we shall discuss some properties of the optimal Bernoulli schedule. In particular, light- and heavy-traffic asymptotes of $\boldsymbol{q}^*$ will be discussed; they are based on light-traffic asymptotes of the mean waiting times and the stability condition (6), respectively. Moreover, based on the μc-rule, a partial solution to the optimization problem will be presented; that is, explicit values of q_i^* for some particular queues i will be given.

3. Light-traffic and heavy-traffic properties; partial solution

The most general exact results obtained for polling systems are the formulations of the so-called pseudo-conservation laws (cf. [8]); a pseudo-conservation law is an exact expression for a specific weighted sum of the mean waiting times at the various queues. For a cyclic polling system with Bernoulli schedules at all queues the pseudo-conservation law reads (cf. [6]):

$$\sum_{i=1}^{s} \rho_i \left[1 - a_i(1-q_i)\frac{\sigma_1 \rho}{1-\rho}\right] EW_i = \frac{\rho^2}{1-\rho}\frac{\beta_2}{2\beta_1} + \rho\frac{\sigma_2}{2\sigma_1} + \frac{\sigma_1}{1-\rho}\sum_{i=1}^{s}\rho_i^2(1-q_i) + \frac{\sigma_1}{2(1-\rho)}\left[\rho^2 - \sum_{i=1}^{s}\rho_i^2\right]. \tag{9}$$

Note that in the case $c_i \equiv \eta_i$ $(i = 1, \ldots, s)$, $C(\boldsymbol{q})$ is proportional to the mean amount of unfinished work in the system. The latter can be shown to be minimal when all queues are served exhaustively (cf. [22]), i.e. $q_i^* = 1$ for $i = 1, \ldots, s$. Moreover, it follows from (9) that if $\sigma_1 = 0$, then $C(\boldsymbol{q})$ is independent of $\boldsymbol{q}$, so that every feasible value of $\boldsymbol{q}$ is optimal.

3.1. Light-traffic properties

An elegant tool to get an insight into the light-traffic behaviour of the system is the use of the PSA. The algorithm relies on the property that the state probabilities can be expressed as power-series in ρ, for all values of ρ for which the stability condition (6) is satisfied. The coefficients of these state probabilities can be determined by means of an iteration scheme (cf. [2,5]). Once these coefficients have been determined, those of the mean queue lengths can be obtained as well (cf. [2]). Then, the coefficients of the power-series expansions of the mean waiting times follow with the aid of Little's formula.

Now, *algebraically* determining the first few terms of the power-series expansions of the state probabilities and then, of the mean waiting times, gives an insight into the light-traffic behaviour of the system and moreover, provides useful information about the optimization of Bernoulli schedules in light-traffic systems. The algebraic computation of the first few terms of the power-series expansions of the mean waiting times according to the computation schemes (cf. [2,5]) is tedious, but straightforward. To this end, we used an algebraic formula manipulation computer program, yielding the following light-traffic asymptotes of the mean waiting times:

(a) if $\sigma_1 = 0$, then for $i = 1, \ldots, s$,

$$EW_i = \rho\frac{\beta_2}{2\beta_1} + \rho^2\left[\eta_i\frac{\beta_2}{2\beta_1} + \Xi(i) + \Psi(i, \boldsymbol{q})\right] + O(\rho^3), \quad (\rho \downarrow 0), \tag{10}$$

with

$$\Xi(i) := \sum_{j=1}^{s-1}\eta_{i+j}\sum_{k=0}^{j-1}\eta_{i+k}\frac{\beta_{i+k}^{(2)}}{\beta_{i+k}^{(1)}}, \qquad \Psi(i, \boldsymbol{q}) := \sum_{j=1}^{s} q_j\eta_j^2\frac{\beta_j^{(2)}}{\beta_j^{(1)}} - q_i\eta_i\frac{\beta_i^{(2)}}{\beta_i^{(1)}}; \tag{11}$$

(b) if $\sigma_1 > 0$, then for $i = 1, \ldots, s$,

$$EW_i = \frac{\sigma_2}{2\sigma_1} + \rho\left[\frac{\beta_2}{2\beta_1} + \left(\sigma_1 - \frac{\sigma_2}{2\sigma_1}\right)(1-\eta_i) + \Phi(i) + \Omega(i, \boldsymbol{q})\right] + O(\rho^2), \quad (\rho \downarrow 0), \tag{12}$$

with

$$\Phi(i) := \sum_{\substack{j=1 \\ j\neq i}}^{s} \sum_{k=i+1}^{j} \eta_j \frac{\sigma_k^{(2)} - \sigma_k^{(1)^2}}{\sigma_1}, \qquad \Omega(i, \boldsymbol{q}) := (1-q_i)\left(\eta_i\sigma_1 + \tfrac{1}{2}a_i\sigma_2\right). \tag{13}$$

Here, the indices, say i, should be replaced by i MOD s if $i > s$.

Remarks. The PSA is only applicable to systems which can be modeled as quasi birth-death processes. General probability distributions for the service times and the switch-over times are approximated by Coxian distributions. However, the capacity of the symbolic manipulator has restricted the computation to systems with 3 queues, using 2-phase Coxian distributions. The so-obtained coefficients in (10) and (12) have been checked numerically for numerous examples with more than 3 queues and more than 2 phases, and were found to be valid in all considered cases. Moreover, one may verify that the coefficients in (10) and (12) are in agreement with the pseudo-conservation law (9). The coefficient of ρ^2 in (10) consists of three parts. The term $\Xi(\cdot)$ reflects the influence of the order in which the server visits the queues and the term $\Psi(\cdot, \cdot)$ depends on the service discipline at the queues. The coefficient of ρ in (12) consists of four parts. The term $\Phi(\cdot)$ depends on the individual switch-over time distributions and is independent of the service disciplines; the term $\Omega(\cdot, \cdot)$ depends on the switch-over time distributions only through the total switch-over time distribution per cycle, and it depends on the service disciplines at the queues.

For models with zero and non-zero switch-over times light-traffic asymptotes of $\boldsymbol{q}^*$ can be derived from (10) and (12) respectively. Let us first consider the case $\sigma_1 = 0$. Omitting the terms which do not depend on $\boldsymbol{q}$ in (10), it follows that the coefficient of ρ^2 of the cost function (7) is

$$\sum_{i=1}^{s} q_i\eta_i(\eta_i - c_i)\frac{\beta_i^{(2)}}{\beta_i^{(1)}}, \tag{14}$$

yielding the following light-traffic asymptotes of $\boldsymbol{q}^*$:
(a) if $\sigma_1 = 0$, then for $i = 1, \ldots, s$,
(i) if $c_i > \eta_i$, then

$$\lim_{\rho \downarrow 0} q_i^* = 1, \tag{15}$$

(ii) if $c_i < \eta_i$, then

$$\lim_{\rho \downarrow 0} q_i^* = 0. \tag{16}$$

The case $c_i = \eta_i$ is not covered by (15) and (16); apparently, the light-traffic limit of q_i^* is determined by higher order terms of the power-series expansions (10). Let us now consider the case $\sigma_1 > 0$. Then (7), (12) and (13) yield the following light-traffic limit of $\boldsymbol{q}^*$:
(b) if $\sigma_1 > 0$, then for $i = 1, \ldots, s$,

$$\lim_{\rho \downarrow 0} q_i^* = 1. \tag{17}$$

Remarks. The limits are defined such that the total arrival rate tends to zero, while the ratios between the individual arrival rates remain fixed.

If σ_1 becomes very small, the difference in the cost associated with $q_i = 0$ and $q_i = 1$ tends to zero, in light traffic. Numerical experience suggests that the individual mean waiting times are smooth functions of $\boldsymbol{q}$. Therefore, one would expect that (15), (16) and (17) not only hold in the limiting case $\rho \downarrow 0$, but remain valid when $0 \le \rho < \epsilon$, for some ϵ small enough. In fact, cases were found with $\sigma_1 > 0$ small and $c_i < \eta_i$, where there exist positive numbers $\rho^{(0)}$, $\rho^{(1)}$ and $\rho^{(2)}$, satisfying $0 < \rho^{(1)} < \rho^{(2)} < \rho^{(0)}$, such that $q_i^* = 1$ for $0 \le \rho \le \rho^{(1)}$, q_i^* decreases from one to zero as ρ increases from $\rho^{(1)}$ to $\rho^{(2)}$ and $q_i^* = 0$ for $\rho^{(2)} \le \rho \le \rho^{(0)}$, cf. e.g. Fig. 3 in Section 5. The fact that $\boldsymbol{q}$ appears in the ρ-term in (12) and does *not* appear in the ρ-term in (10) can be explained as follows. As described in [3], the ρ^k-terms in (10) and (12) correspond to states of the system in which at most k customers are present in the system, and using PASTA, to situations in which an arriving customer finds at most k customers present in the system upon arrival, $k = 0, 1, \ldots$. Now, consider a marked customer C_A arriving at queue i, while there is only one customer, C_B, present in the system, which is residing at queue i. Moreover, suppose that there are no arrivals during the waiting time of C_A (higher order effect). Then the waiting time of C_A is equal to the residual sojourn time of C_B *plus, with probability* $1 - q_i$, a complete cycle of the server along the queues. Hence, if $\sigma_1 > 0$, then the waiting time of C_A clearly depends on q_i. If $\sigma_1 = 0$, the extra cycle time vanishes, because of the assumption that no other customers arrive during the waiting time of C_A and hence, the waiting time of C_A does *not* depend on q_i $(i = 1, \ldots, s)$.

3.2. Heavy-traffic properties

Denote the heavy-traffic residue of the mean waiting time at queue i by

$$\omega_i := \lim_{\chi \uparrow 1} (1 - \chi) EW_i, \quad i = 1, \ldots, s. \tag{18}$$

These limits are defined in such a way that the total arrival rate to the system increases to a value at which one or more queues become instable, while the proportions between the arrival rates remain fixed. When the number of queues is not too large and the parameters of a system are not too asymmetrical, it is possible to obtain accurate data for performance measures even for high occupancy of the system, i.e., χ close to one, by means of the PSA. From those results the heavy-traffic residues in (18) can be estimated. An important general observation, which is supported by results in [12], is the following:

for $i = 1, \ldots, s$, it holds that the residue ω_i is positive if

$$a_i(1 - q_i) = \max_{j = 1, \ldots, s} \{a_j(1 - q_j)\}; \tag{19}$$

otherwise, $\omega_i = 0$ and EW_i possesses a finite limit as $\chi \uparrow 1$.

In general, there is no closed-form expression for ω_i, except for a few exceptional cases. As an example of such a special case, consider a system with exhaustive service at all queues, i.e., $q_i = 1$, $i = 1, \ldots, s$. For this system one can obtain explicit expressions for heavy-traffic residues of the mean waiting times from the set of equations in [29, Chapter 4]:

if $q_i = 1$, $i = 1, \ldots, s$, then the heavy-traffic residue of the mean waiting time at queue i is given by

$$\omega_i = \frac{1-\eta_i}{\sum_{j=1}^{s} \eta_j(1-\eta_j)} \frac{\beta_2}{2\beta_1} + \tfrac{1}{2}\sigma_1(1-\eta_i), \quad i = 1, \ldots, s. \tag{20}$$

Moreover, if $\omega_i = 0$, the finite limit for the mean waiting time at queue i does not admit a closed-form expression, except for a few special cases, e.g., when (19) is satisfied for all queues, except for one queue.

If $\sigma_1 > 0$, the stability condition (6) can be reformulated as

$$q_i > 1 - \frac{1-\rho}{\sigma_1 a_i \rho}, \quad i = 1, \ldots, s. \tag{21}$$

Hence, (21) induces a lower bound on the set of possible values of q_i^*, $i = 1, \ldots, s$. Moreover, as the right-hand side of (21) tends to one as $\rho \uparrow 1$, we have the following heavy-traffic asymptote for q_i^*:

if $\sigma_1 > 0$, then

$$\lim_{\rho \uparrow 1} q_i^* = 1, \quad i = 1, \ldots, s. \tag{22}$$

Remark. If $\sigma_1 = 0$, then the stability condition reads $\rho < 1$ and hence, $\boldsymbol{q}$ has no influence on the stability of the system. Thus, unlike in the case $\sigma_1 > 0$, the ergodicity condition (6) does *not* imply that q_i^* tends to 1, for $i = 1, \ldots, s$, as $\rho \uparrow 1$. In fact, cases have been found in which the heavy-traffic asymptote of q_i^* lies in the interior of [0,1], cf., e.g., Fig. 4 in Section 5.

3.3. Partial solution

The following observations may reduce the dimension of the optimization problem:

(a) if $c_i/\eta_i = \max_{j=1,\ldots,s} \{c_j/\eta_j\}$, then $q_i^* = 1$, (23)

(b) if $\sigma_1 = 0$, then

if $c_i/\eta_i = \min_{j=1,\ldots,s} \{c_j/\eta_j\}$, then $q_i^* = 0$. (24)

Properties (23) and (24) are supported by the μc-rule for priority systems. For systems with zero switch-over times, in order to minimize a weighted sum of the mean waiting times, the μc-rule gives priorities to the queues in increasing order of the values of c_i/η_i (cf. [24]). Hence, if a queue i, for which $c_i/\eta_i = \max_{j=1,\ldots,s} \{c_j/\eta_j\}$, is non-empty upon a service completion epoch at queue i, then according to the μc-rule, it is not optimal for the server to proceed to

another queue; and if the server has just completed a service at queue k, for which $c_k/\eta_k = \min_{j=1,\ldots,s}\{c_j/\eta_j\}$, then according to the μc-rule, the server should depart from queue k to check if customers are waiting for service at other queues. When the switch-over times increase, the optimal values of the Bernoulli parameters tend to increase, because the server should serve more jobs at each visit to compensate for the loss of its availability due to the switches. This implies that when $\sigma_1 > 0$ a queue i with c_i/η_i maximal should still be served exhaustively, while a queue i with c_i/η_i minimal may require a positive q_i (cf. also (21)).

Remarks. Properties (23) and (24) might decrease the dimension of the numerical solution of the optimization problem considerably. The number of components of $\boldsymbol{q}^*$ which are determined by (23) and (24) varies from 1 to s, depending on the system parameters. In the case $c_i \equiv \eta_i$ $(i = 1,\ldots,s)$, $\sigma_1 > 0$, it follows from (23) that $q_i^* \equiv 1$ $(i = 1,\ldots,s)$, which is in agreement with the results in [22]. Moreover, in the case $c_i \equiv \eta_i$ $(i = 1,\ldots,s)$, $\sigma_1 = 0$, (23) and (24) are in agreement with the fact that $C(\boldsymbol{q})$ does not depend on $\boldsymbol{q}$, so that every feasible value of $\boldsymbol{q}$ is optimal (cf. also the remarks below (9)).

4. Numerical optimization

In the previous section properties of optimal Bernoulli schedules have been presented. In general, however, the optimization problem is not analytically solvable. Therefore, a numerical approach to compute the optimal Bernoulli schedule accurately will be discussed in this section. The approach is based on the use of the PSA, which, in principle, can compute $C(\boldsymbol{q})$ within any level of accuracy, for any feasible value of $\boldsymbol{q}$. Combining the PSA with some (local) optimization procedure may lead to an accurate determination of $\boldsymbol{q}^*$. In optimization theory numerous algorithms to find (local) optima are available, using information about the characteristics of the function to be optimized. The optimization techniques can be classified into two classes, depending on whether or not they use derivatives: the direct search methods, which do not use derivatives, and the gradient methods, which are generally more efficient. The reader is referred to [26, Chapter 6] for an overview of the various optimization techniques. As the time requirements of the use of the PSA for evaluation purposes may be considerable, efficiency of the optimization procedure is of great importance (cf. [5] for a discussion on the complexity of the PSA). The memory requirements restrict the use of the PSA to queueing systems with a fairly small number of queues and moreover, limit the number of terms in the power-series expansions that can be computed. Hence, restrictions on the available computation time and memory space may affect the accuracy of the mean waiting time approximations which, in turn, may affect the accuracy of the PSA-based optimal Bernoulli schedule. Because little is known about the cost function $C(\boldsymbol{q})$ as a function of $\boldsymbol{q}$, one is forced to rely on numerical experience, which suggests that $C(\boldsymbol{q})$ is smooth as function of $\boldsymbol{q}$, so that gradient methods may be applied for optimization. Because of efficiency and memory considerations, the authors propose to use the so-called conjugate gradient method (cf. [26, Section 6.3]), with the obvious modification that the Bernoulli parameters are restricted to the interval [0,1]. Here, the partial derivatives of the cost function (7) with respect to the Bernoulli parameters are estimated on the basis of

finite differences; for a feasible $\boldsymbol{q}$ the partial derivative of the cost function w.r.t. q_j $(j = 1, \ldots, s)$ is estimated according to the following formula:

$$\frac{C(\boldsymbol{q} + h\mathrm{e}_j) - C(\boldsymbol{q} - h\mathrm{e}_j)}{2h}, \tag{25}$$

where e_j is the jth unit vector in $\mathbb{R}^s$ $(j = 1, \ldots, s)$ and where h is a suitably chosen step size. We refer to [14] for an extensive discussion of implementation of gradient methods based on finite difference approximations. In practice, the step sizes used to estimate the finite differences according to (25) and the step sizes in the line searches are adapted to the accuracy of the computed cost. The latter can only be estimated on the basis of differences in the last few computed terms of the power-series, cf. [4,5] for a discussion of the accuracy of the PSA.

Remark. The PSA is also applicable to a variety of non-cyclic routing mechanisms (cf. [5]). Hence, the proposed optimization method can also be applied to systems with those routing mechanisms.

This section is concluded with plots of $C(\boldsymbol{q})$ for a few numerical examples. First, to illustrate that $C(\boldsymbol{q})$ is a smooth function of $\boldsymbol{q}$, consider the model with the following set of parameters: $s = 3$; $\beta^{(1)} = (1.0, 2.0, 3.0)$; $\sigma^{(1)} = (0.15, 0.15, 0.15)$; all service times and switch-over times are exponentially distributed; $a = (1/6, 1/6, 1/6)$; $c = (0.40, 0.25, 0.35)$; $\rho = 0.8$. For this model, $q_1^* = 1$, in agreement with (23). A typical plot of $C(\boldsymbol{q})$ for $q_1 = 1$ as function of q_2, q_3 is shown in Fig. 1.

In general, $C(\boldsymbol{q})$ is not a convex function of $\boldsymbol{q}$, so that local optimality does not guarantee global optimality of an optimum found by means of a local optimization procedure. Nevertheless, it has been supported by numerous numerical experiments with different initial schedules that $C(\boldsymbol{q})$ has no alternative local minima. To illustrate the latter, consider the following model: $s = 2$; $\beta^{(1)} = (0.05, 0.30)$; $\sigma^{(1)} = (0.001, 0.001)$; all service times and switch-over times are exponentially distributed; $a = (20/7, 20/7)$; $c = (0.1, 0.9)$; $\rho = 0.8$. Fig. 2 shows a typical plot of $C(\boldsymbol{q})$ for $q_1, q_2 \in [0,1]$. Note that, in agreement with (23), in this example we have $q_2^* = 1$.

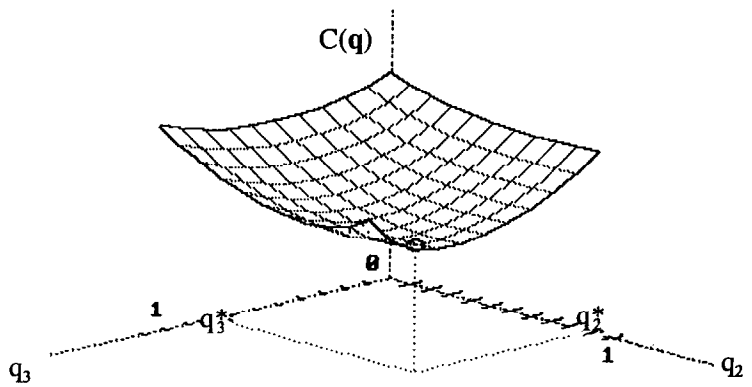


Fig. 1. Typical plot of $C(\boldsymbol{q})$.

5. Influence of system parameters on the optimal schedule

In this section, the influence of system parameters on the optimal Bernoulli schedule is discussed. The optimal schedules have been obtained by means of the numerical optimization technique discussed in the previous section.

5.1. Offered load

Optimal Bernoulli schedules have been computed for various values of the offered load; here, the load is varied in such a way that the ratios between the arrival rates remain fixed. Consider the model with the following set of parameters: $s = 3$; $\beta^{(1)} = (0.5, 1.0, 1.5)$; $\sigma^{(1)} = (r, r, r)$; all service times and switch-over times are exponentially distributed; $a = (1/3, 1/3, 1/3)$; $c = (10/55, 15/55, 30/55)$. For this model, $q_1^* = q_3^* = 1$, in agreement with (23). Fig. 3 shows the values of q_2^* as function of ρ for varying values of r. Note that in the case $r = 0$ we have $q_2^* = 0$, in agreement with (24).

As discussed in Section 3, the characteristics of the optimal schedule as function of ρ in the case $\sigma_1 = 0$ may differ from those in the case $\sigma_1 > 0$. To illustrate this, optimal schedules for varying values of the offered load have been computed for the model with the following set of parameters: $s = 3$; $\beta^{(1)} = (0.5, 1.0, 1.5)$; all service times at the queues 1 and 3 are exponentially distributed and the service times at queue 2 are 2-phase Coxian with squared coefficient of variation 4; $\sigma_1 = 0$; $a = (1/3, 1/3, 1/3)$. Two different cost functions are considered: $c_1 = (0.6, 0.3, 0.1)$ and $c_2 = (0.50, 0.35, 0.15)$. In agreement with (23) and (24), in both cases we have $q_1^* = 1$ and $q_3^* = 0$. Fig. 4 shows q_2^* as function of the offered load for both cost functions.

Remark. Figs. 3 and 4 confirm the validity of the light- and heavy-traffic asymptotes, cf. (15), (16), (17) and (22). Fig. 4 illustrates the fact that in the case $\sigma_1 = 0$ some of the components of the heavy-traffic asymptote of $\boldsymbol{q}^*$ may have values in the interior of the interval [0,1], as remarked at the end of Section 3.

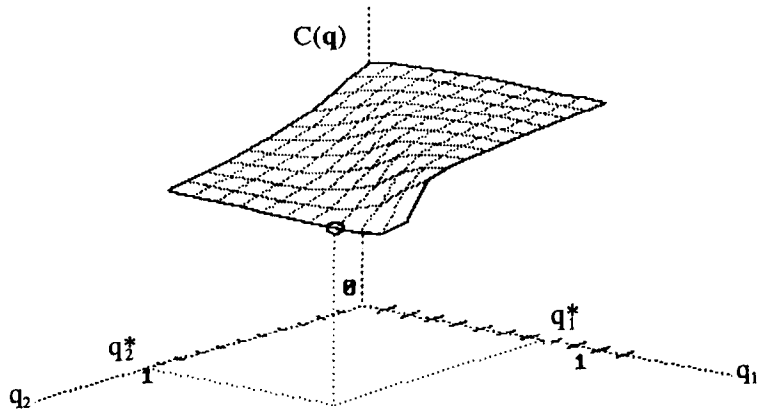


Fig. 2. Example of a non-convex cost function.

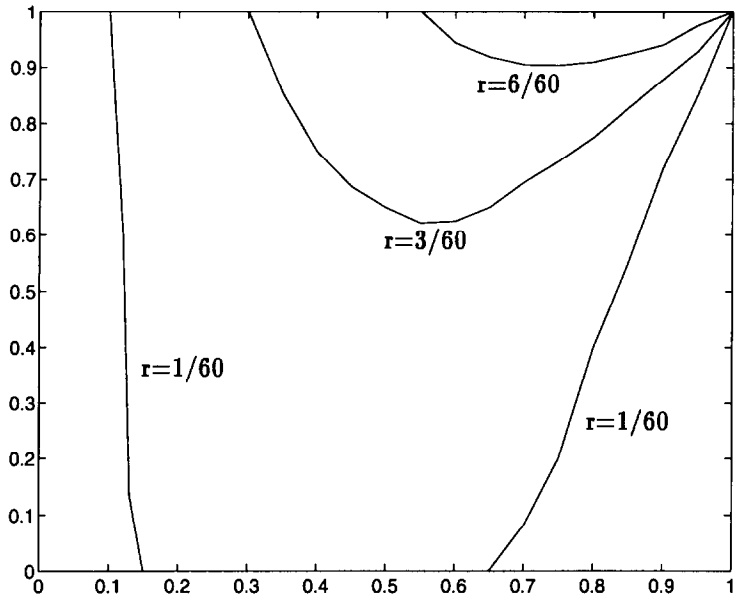


Fig. 3. The optimal value of q_2^* as function of the offered load; $\sigma_1 > 0$.

We shall now discuss the influence of system parameters which are independent of the offered load: the service time distributions and the switch-over time distributions. In general, the mean waiting times depend on higher moments of the service time and switch-over time distributions. However, the mean waiting times depend on the service time and switch-over

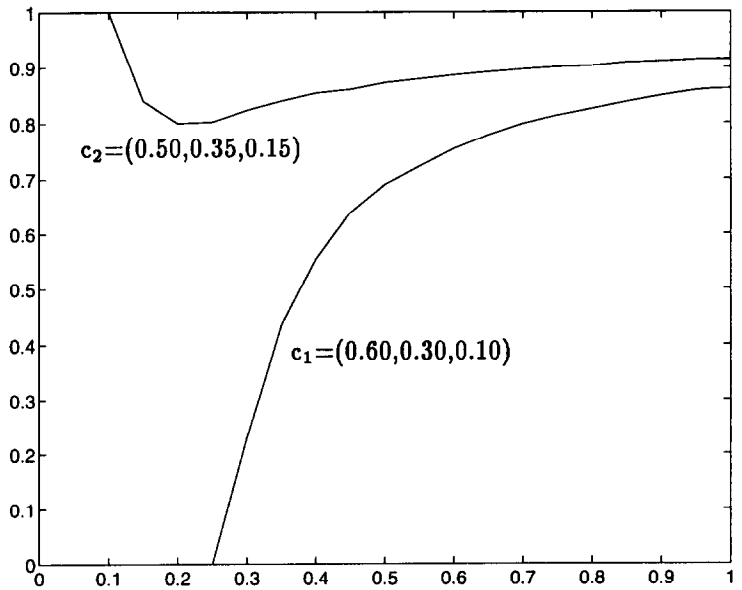


Fig. 4. The optimal values of q_2^* as function of the offered load; $\sigma_1 = 0$.

Table 1
The optimal schedule for different values of $\beta^{(2)}$ for β_2 fixed

$\beta^{(2)}$	$r = 0.01$		$r = 0.05$		$r = 0.50$	
	(q_2^*, q_3^*)	$C(\cdot)$	(q_2^*, q_3^*)	$C(\cdot)$	(q_2^*, q_3^*)	$C(\cdot)$
(2.0, 8.0, 20.0)	(0.26, 0.00)	8.22	(0.59, 0.22)	9.06	(1.00, 0.91)	13.48
(2.0, 10.0, 18.0)	(0.25, 0.00)	8.23	(0.57, 0.21)	9.08	(1.00, 0.90)	13.52
(4.0, 8.0, 18.0)	(0.27, 0.00)	8.22	(0.60, 0.22)	9.06	(1.00, 0.91)	13.46
(1.5, 6.0, 22.5)	(0.28, 0.00)	8.21	(0.62, 0.23)	9.05	(1.00, 0.90)	13.44
(1.5, 15.0, 13.5)	(0.25, 0.00)	8.24	(0.54, 0.21)	9.09	(1.00, 0.90)	13.61

time distributions primarily through the first two moments of the service times and the switch-over times, as illustrated in [5, Table 1] and [4, Section 5], respectively. Therefore, we only consider the influence of the first two moments of the service and switch-over times on the optimal schedule.

The observations are based on numerous numerical experiments and have been confirmed in all cases considered; we emphasize that the observations provide rough guidelines to simplify the problem (e.g. to reduce the size of the state space) and we do not pretend that these observations remain valid for all values of the model parameters.

5.2. Service time distributions

Numerical experience has indicated that the optimal schedules and in particular, the optimal cost, seem to depend primarily on the second moments of the service time distributions through β_2, rather than on the second moments of the individual service time distributions, $\beta^{(2)}$. Note that this observation is supported by the pseudo-conservation law (9). To illustrate this, the value of $\boldsymbol{q}^*$ has been computed for combinations of the individual second moments of the service time distributions, $\beta^{(2)}$, where the second moment of the service time distribution of an arbitrary customer, β_2, is kept fixed. Table 1 shows the value of $\boldsymbol{q}^*$ for different combinations $\beta^{(2)}$, which are constructed in such a way that $\beta_2 = 10$ in all considered cases. The other parameters are taken to be: $s = 3$; $\beta^{(1)} = (1.0, 2.0, 3.0)$; all service times are 2-phase Coxian distributed; $\sigma^{(1)} = (r, r, r)$; all switch-over times are exponentially distributed; $a = (1/6, 1/6, 1/6)$; $\rho = 0.8$; $c = (0.40, 0.25, 0.35)$. In agreement with (23) we have $q_1^* = 1$. Table 1 shows the results for $r = 0.01$, 0.05 and 0.50, respectively. The optimal schedule is equal to (1, 1, 1) for r large enough, and hence, the observation remains valid for large values of r.

The influence of the second moment of the service time distribution of an arbitrary customer, β_2, on the optimal schedule seems to be rather unpredictable; in fact, in some cases components of the optimal schedule decrease for increasing values of β_2, whereas in other cases components increase for increasing β_2. Moreover, the optimal cost increases when β_2 is increased. To illustrate this, the optimal schedules have been computed for different values of β_2 for two different models. The first model is determined by the following set of parameters: $s = 3$; $\beta^{(1)} = (1.0, 2.0, 3.0)$; all service times at queues 1 and 2 are exponentially distributed, and the service times at queue 3 are 2-phase Coxian distributed with squared coefficient of variation α; $\sigma^{(1)} = (r, r, r)$; all switch-over times are exponentially distributed; $a = (1/6, 1/6,$

Table 2
Increasing effect of β_2 on the optimal schedule

β_2	$r = 0.01$		$r = 0.05$	
	(q_2^*, q_3^*)	$C(\cdot)$	(q_2^*, q_3^*)	$C(\cdot)$
7.08	(0.09, 0.00)	5.90	(0.52, 0.15)	6.62
7.83	(0.15, 0.00)	6.50	(0.54, 0.17)	7.25
9.33	(0.23, 0.00)	7.69	(0.58, 0.20)	8.51
12.33	(0.33, 0.00)	10.07	(0.64, 0.27)	11.01
18.33	(0.43, 0.00)	14.81	(0.74, 0.31)	15.95

1/6); $\rho = 0.8$; $c = (0.40, 0.25, 0.35)$. In agreement with (23) we have $q_1^* = 1$. Table 2 shows the optimal schedules for $\alpha = 0.25, 0.50, 1.00, 2.00$ and 4.00 and for $r = 0.01$ and 0.05, respectively.

For the second model the parameters are: $s = 3$; $\beta^{(1)} = (0.5, 1.0, 1.5)$; all service times at queues 1 and 3 are exponentially distributed, and the service times at queue 2 are 2-phase Coxian distributed with squared coefficient of variation α; $\sigma^{(1)} = (r, r, r)$; all switch-over times are exponentially distributed; $a = (1/3, 1/3, 1/3)$; $\rho = 0.8$; $c = (0.60, 0.30, 0.10)$. In agreement with (20) we have $q_1^* = 1$. Table 3 shows the optimal schedules for $\alpha = 1, 2, 3$ and 4, and $r = 0.01$ and 0.05, respectively.

5.3. Switch-over time distributions

In general, the mean waiting times may depend strongly on the switch-over time distributions. However, as discussed in [4, Section 5], the mean waiting times depend on the switch-over time distributions mainly through the first two moments of the *total* switch-over times during one cycle of the server along the queues. Consequently, as for the first moments of the switch-over times, the optimal Bernoulli schedule depends on $\sigma^{(1)}$ primarily through the first moment σ_1 of the total switch-over time per cycle and moreover, increasing the mean switch-over time per cycle generally leads to an increase of the components of $\boldsymbol{q}^*$. To illustrate this, Table 4 shows the optimal Bernoulli schedules for varying combinations $\sigma^{(1)}$ of the individual mean switch-over times; the total switch-over times consist of three i.i.d. exponential phases, each with mean r and hence, are Erlangian-3 distributed with mean $\sigma_1 = 3r$ in all considered cases. The other model parameters are: $s = 3$; $\beta^{(1)} = (1.0, 2.0, 3.0)$; all service times are exponentially distributed; $a = (1/6, 1/6, 1/6)$; $\rho = 0.8$; $c = (0.40, 0.25, 0.35)$. In agreement

Table 3
Decreasing effect of β_2 on the optimal schedule

β_2	$r = 0.01$		$r = 0.05$	
	(q_2^*, q_3^*)	$C(\cdot)$	(q_2^*, q_3^*)	$C(\cdot)$
2.33	(0.99, 0.00)	2.45	(1.00, 0.00)	2.81
2.67	(0.91, 0.00)	2.80	(0.95, 0.00)	3.18
3.00	(0.86, 0.00)	3.13	(0.91, 0.00)	3.54
3.33	(0.83, 0.00)	3.45	(0.87, 0.00)	3.89

Table 4
The optimal schedule for different values of $\sigma^{(1)}$, with σ_1 (and σ_2) fixed

$\sigma^{(1)}$	$r = 0.05$		$r = 0.15$		$r = 0.50$	
	(q_2^*, q_3^*)	$C(\cdot)$	(q_2^*, q_3^*)	$C(\cdot)$	(q_2^*, q_3^*)	$C(\cdot)$
(r, r, r)	(0.58, 0.21)	8.51	(0.91, 0.66)	9.87	(1.00, 0.91)	12.83
$(3r, 0, 0)$	(0.58, 0.20)	8.51	(0.91, 0.66)	9.87	(1.00, 0.91)	12.82
$(0, 3r, 0)$	(0.58, 0.20)	8.51	(0.91, 0.66)	9.87	(1.00, 0.91)	12.85
$(0, 0, 3r)$	(0.58, 0.20)	8.51	(0.91, 0.66)	9.87	(1.00, 0.91)	12.83

Table 5
The optimal schedule for different values of σ_2 with σ_1 fixed

σ_2/σ_1^2	$r = 0.15$		$r = 0.50$		$r = 1.50$	
	(q_2^*, q_3^*)	$C(\cdot)$	(q_2^*, q_3^*)	$C(\cdot)$	(q_2^*, q_3^*)	$C(\cdot)$
1.25	(0.58, 0.20)	8.50	(0.92, 0.69)	10.02	(1.00, 0.91)	12.76
3.00	(0.58, 0.21)	8.61	(0.92, 0.69)	10.42	(1.00, 0.91)	14.03
5.00	(0.58, 0.21)	8.74	(0.93, 0.69)	10.88	(1.00, 0.90)	15.46
11.00	(0.58, 0.23)	9.12	(0.93, 0.69)	12.22	(1.00, 0.90)	19.67
21.00	(0.60, 0.23)	9.76	(0.94, 0.70)	14.43	(1.00, 0.91)	26.11

with (23) we have $q_1^* = 1$. Table 4 shows the results for $r = 0.05$, $r = 0.15$ and $r = 0.5$, respectively.

As far as the second moments of the switch-over time distributions are concerned, we found out that their effect on the optimal schedule is generally negligible and moreover, that the optimal cost increases when σ_2 is increased. To illustrate this, consider the same model as in Table 4 with $\sigma^{(1)} = (r, 0, 0)$. The optimal schedules have been computed for vaying values of α, the squared coefficients of variation of the switch-over times. Table 5 shows the results for $r = 0.15$, 0.50, 1.50 and $\alpha = 0.25$, 2.00, 4.00, 10.00 and 20.00.

As noted before, below Table 1, in the cases considered in Tables 4 and 5, the optimal schedule is equal to (1, 1, 1) for r large enough, so that the observations remain valid for large values of r. The foregoing suggests that replacing constant switch-over times by Erlang-distributed switch-over times may yield good approximations for the optimal schedules for models with constant switch-over times.

6. Approximation

In Section 4 we have discussed a numerical approach to achieve an accurate approximation for the optimal Bernoulli schedule, based on the use of the PSA. The main disadvantage of this approach is the fact that the time and memory requirements increase exponentially with increasing number of queues and hence, its use is restricted to rather small systems. For this reason, in this section we will propose a simple and fast approach to approximate the optimal Bernoulli schedule, which requires negligible computation time and memory space and is

therefore applicable to fairly large systems. This approximate optimal Bernoulli schedule may serve as a starting point for a more accurate optimization procedure based on the use of the PSA. The approach is based on a simple mean waiting time approximation (instead of on the use of the PSA); combined with the optimization procedure discussed in Section 4 it yields an approximation for the optimal Bernoulli schedule. Consider the following mean waiting time approximation (cf. [30, Section 6.7]):

$$EW_i \approx \frac{(1-\rho+\rho_i)-q_i\rho_i(2-\rho)}{1-\rho-\lambda_i(1-q_i)\sigma_1}x; \tag{26}$$

the quantity x is determined by substituting (26) into the pseudo-conservation law (9). The approximation is a trivial extension of the pseudo-conservation law-based mean waiting time approximation proposed in [15] for mixtures of 1-limited and exhaustive (and gated) service disciplines. The latter approach relies on the observation of Everitt [11], who states that the cycle time of queue i, i.e., the length of the time interval between two successive arrivals of the server to queue i, has approximately the same second moment for all $i = 1, \ldots, s$. Although the approximation is generally not very accurate for evaluation purposes (cf. numerical results in [30, Section 6.7]), combined with the optimization procedure as discussed in Section 4 it turns out to give satisfying results for optimization purposes.

Remarks. The mean waiting time approximation (26) depends on the individual service time distributions only through β_1 and β_2, and similarly, depends on the switch-over time distributions only through σ_1 and σ_2 (cf. (9)). As discussed in Section 5, the optimal schedules and, in particular, the cost of the optimal schedule, are fairly insensitive to the individual service time and switch-over time distributions. As noted in Section 3, in the case $\sigma_1 = 0$ the heavy-traffic asymptote of $\boldsymbol{q}^*$ is generally unknown. Therefore, the heavy-traffic asymptote of approximated optimum based on (26) may differ from the exact heavy-traffic asymptote. Hence, the accuracy of the approximated optimum and in particular, the cost belonging to the approximated optimum may become poor when $\sigma_1 = 0$ and the offered load approaches 1. Moreover, in spite of the fact that for $\sigma_1 > 0$, the heavy-traffic limits of the approximated optima and the actual optima are both equal to 1, the accuracy of the approximated optimum and the cost belonging to this optimum, may become poor, particularly in cases in which $\sigma_1 \approx 0$ and $\rho \approx 1$. This is due to the fact that for $\sigma_1 \approx 0$ the lower bound of the values of the components of $\boldsymbol{q}$ tends to 1 very slowly, so that for values of ρ close to 1 the lower bound does not force the values of $\boldsymbol{q}$ into a narrow interval.

Fig. 5 illustrates the behaviour of the approximated optimum, $\boldsymbol{q}^*(\text{app})$, as function of the offered load ρ, compared with the optimum $\boldsymbol{q}^*$, obtained by means of the numerical optimization technique discussed in Section 4. The parameters are: $s = 3$; $\beta^{(1)} = (0.5, 1.0, 1.5)$; $\sigma^{(1)} = (0.05, 0.05, 0.05)$; all service times and switch-over times are exponentially distributed; $a = (1/3, 1/3, 1/3)$; $c = (0.40, 0.25, 0.35)$. In agreement with (23) we have $q_1^* = 1$.

The quality of the approximation can best be measured by the relative difference between the exact cost belonging to $\boldsymbol{q}^*(\text{app})$, $C(\boldsymbol{q}^*(\text{app}))$, and the cost of the optimum, $C(\boldsymbol{q}^*)$, rather

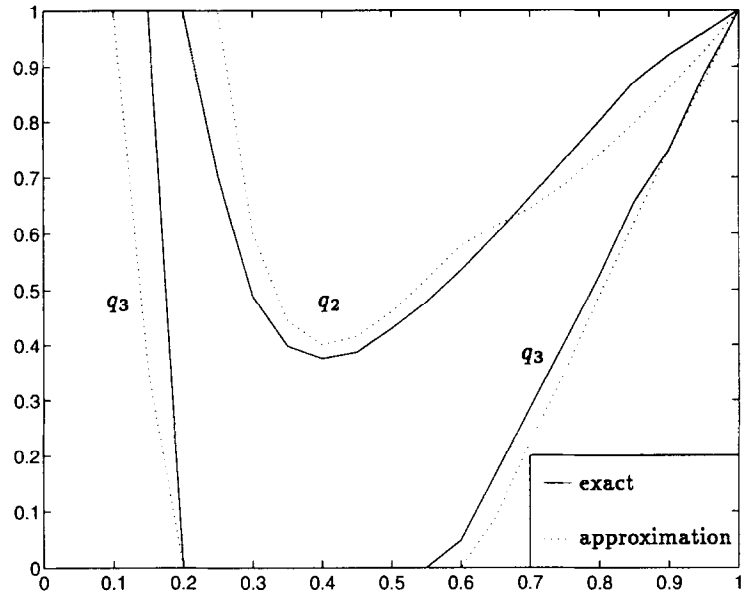


Fig. 5. Behaviour of $q^*(\text{app})$ and q^* as function of ρ.

than by the relative accuracy of $q^*(\text{app})$, compared with q^*. Table 6 gives for the same set of parameters the relative error in the cost function, defined by

$$\frac{C(q^*(app)) - C(q^*)}{C(q^*)} \times 100\%. \tag{27}$$

Remark. From Fig. 5 one may observe that the approximated schedule may differ considerably from the optimum when the load is rather small. However, as is also illustrated in Table 6, the relative error in the cost associated with these optima is small; this is due to the fact that for lightly loaded systems the cost function (7) is rather flat as function of q.

In order to check the quality of the approximation for larger systems, consider a 6-queue model with the following set of parameters: $s = 6$; $\beta^{(1)} = (0.60, 0.80, 1.00, 1.20, 1.40, 1.60)$;

Table 6
The accuracy of the cost belonging to the approximated optimum

ρ	$C(q^*(\text{app}))$	$C(q^*)$	err%
0.10	0.236	0.236	0.0
0.15	0.317	0.317	0.0
0.50	1.272	1.272	0.0
0.60	1.836	1.835	0.0
0.70	2.772	2.771	0.0
0.80	4.638	4.635	0.1
0.90	10.231	10.205	0.3
0.95	21.468	21.259	1.0

Table 7
The approximated optima compared with exact optima for a 6-queue model; $r = 0.05$

ρ	q^*_{2-6}(app)	$C(\cdot)$	q^*_{2-6}	$C(\cdot)$	err%
0.50	(0.23, 0.00, 0.00, 0.00, 0.00)	1.34	(0.71, 0.00, 0.00, 0.00, 0.00)	1.34	0.0
0.60	(0.55, 0.00, 0.00, 0.00, 0.00)	1.83	(0.85, 0.00, 0.00, 0.00, 0.00)	1.83	0.1
0.70	(0.84, 0.00, 0.00, 0.00, 0.00)	2.58	(1.00, 0.22, 0.00, 0.00, 0.00)	2.57	0.4
0.80	(1.00, 0.29, 0.00, 0.00, 0.00)	3.98	(1.00, 0.43, 0.00, 0.00, 0.00)	3.97	0.3

Table 8
The approximated optima compared with exact optima for a 6-queue model; $r = 0.10$

ρ	q^*_{2-6}(app)	$C(\cdot)$	q^*_{2-6}	$C(\cdot)$	err%
0.50	(1.00, 0.46, 0.00, 0.00, 0.00)	1.68	(1.00, 0.36, 0.00, 0.00, 0.00)	1.68	0.0
0.60	(1.00, 0.56, 0.00, 0.00, 0.00)	2.26	(1.00, 0.56, 0.00, 0.00, 0.00)	2.26	0.0
0.70	(1.00, 0.69, 0.19, 0.00, 0.00)	3.19	(1.00, 0.71, 0.23, 0.00, 0.00)	3.19	0.0
0.80	(1.00, 0.78, 0.50, 0.28, 0.11)	5.03	(1.00, 0.78, 0.50, 0.28, 0.13)	5.03	0.0

$\sigma_i^{(1)} = r$ for $i = 1, \ldots, 6$; all service times and switch-over time are exponentially distributed; $a_i = 5/33$ for $i = 1, \ldots, 6$; $c = (0.65, 0.07, 0.07, 0.07, 0.07, 0.07)$. In agreement with (23) we have $q_1^* = 1$. Table 7 and Table 8 show the other components of the approximated optimum, q^*_{2-6}(app) and of the exact optimum q^*_{2-6} for varying values of the offered load ρ, for $r = 0.05$ and $r = 0.10$, respectively; the relative error in the cost function (err%) is computed according to (27).

Remark. As illustrated in Tables 7 and 8, for larger systems a main part of the components of the optimal schedule may be equal to 0 or 1. This may be explained by the fact that for larger systems, the offered load to many of the queues is small, so that the waiting times are fairly independent of the values q_i for those queues. Consequently, the cost function (7) is rather flat as function of q_i and hence, may be either increasing or decreasing in q_i over the interval [0,1], so that the optimal value of q_i is equal to either 0 or 1.

In general, the optimal schedule depends on the order in which the queues are visited, whereas the approximated optimum based on (26) does *not* depend on the order in which the queues are placed. However, the differences in the optimal schedules and in particular, in the

Table 9
The approximated optima compared with exact optima for a 6-queue model

ρ	q^*_{2-6}(app)	$C(\boldsymbol{q}^*(\text{app}))$	q^*_{2-6}	$C(\boldsymbol{q}^*)$	err%
0.30	1.00	1.09	(1.00, 1.00, 1.00, 1.00, 1.00)	1.09	0.0
0.40	1.00	1.50	(0.79, 0.77, 0.76, 0.76, 0.76)	1.50	0.0
0.50	0.95	2.08	(0.66, 0.62, 0.59, 0.57, 0.56)	2.06	0.1
0.60	0.82	2.92	(0.66, 0.60, 0.57, 0.55, 0.53)	2.90	0.7
0.70	0.80	4.30	(0.72, 0.67, 0.64, 0.62, 0.60)	4.28	0.5
0.80	0.84	7.04	(0.86, 0.78, 0.73, 0.72, 0.70)	7.01	0.4

cost associated with the optimal schedules, have turned out to be small. To illustrate this, consider the model with the following set of parameters: $s = 6$; $\beta^{(1)} = (0.5, 1.5, 1.5, 1.5, 1.5, 1.5)$; $\sigma^{(1)} = (0.1, 0.1, 0.1, 0.1, 0.1, 0.1)$; all service times and switch-over times are exponentially distributed; $a_i = 0.125$, $i = 1, \ldots, 6$; $c = (0.375, 0.125, 0.125, 0.125, 0.125, 0.125)$. In agreement with (23) we have $q_1^* = 1$. Table 9 shows the approximated optimum $\boldsymbol{q}^*(\text{app})$ and the exact optimum $\boldsymbol{q}^*$ for varying values of the offered load ρ; the relative error in the cost function (err%) is computed according to (27). Note that $q_2^*(\text{app}) = q_i^*(\text{app})$, $i = 3, \ldots, 6$, for all values of ρ; hence, the approximated optimum $\boldsymbol{q}^*(\text{app})$ is denoted by $q_{2-6}^*(\text{app})$.

An alternative approach which is more generally applicable, also for polling models for which no such approximation as (26) is available, is to use the PSA with a small number of terms to find the neighbourhood of the optimal schedule with reduced computational effort, and then proceed with the PSA with more terms to locally improve the optimal schedule. Further, the partial solution (23), (24), may be used to decrease the dimension of the optimization problem.

Acknowledgements

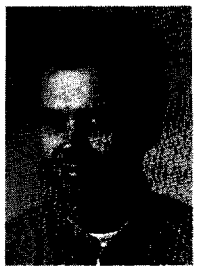
The authors wish to thank Onno Boxma and Jan Weststrate for interesting discussions and useful comments. They also wish to thank the referees for their comments which led to improvement of the presentation of the material.

References

[1] J.P.C. Blanc, On a numerical method for calculating state probabilities for queueing systems with more than one waiting line, *J. Comput. Appl. Math.* **20** (1987) 119–125.
[2] J.P.C. Blanc, A numerical approach to cyclic-service queueing models, *Queueing Systems* **6** (1990) 173–188.
[3] J.P.C. Blanc, Cyclic polling systems: limited service versus Bernoulli schedules, Report FEW 422, Department of Economics, Tilburg University, Tilburg, 1990.
[4] J.P.C. Blanc, The power-series algorithm applied to cyclic polling systems, *Stoch. Models* **7** (1991) 527–545.
[5] J.P.C. Blanc, Performance evaluation of polling systems by means of the power-series algorithm, *Ann. Op. Res.* **35** (1992) 155–186.
[6] O.J. Boxma, Workloads and waiting times in single-server systems with multiple customer classes, *Queueing Systems* **5** (1989) 185–214.
[7] O.J. Boxma, Analysis and optimization of polling systems, *Proc. 13th Int. Teletraffic Congress, Workshop: Queueing, Performance and Control in ATM* (North-Holland, Amsterdam, 1991) 173–183.
[8] O.J. Boxma and W.P. Groenendijk, Pseudo-conservation laws in cyclic-service systems, *J. Appl. Probab.* **24** (1987) 949–964.
[9] O.J. Boxma, H. Levy and J.A. Weststrate, Efficient visit frequencies for polling tables: minimization of waiting cost, *Queueing Systems* **9** (1991) 133–162.
[10] S. Browne and U. Yechiali, Dynamic priority rules for cyclic-type queues, *Adv. Appl. Probab.* **21** (1989) 432–450.
[11] D. Everitt, Simple approximations for token rings, *IEEE Trans. Comm.* **34** (1986) 719–721.
[12] C. Fricker and M.R. Jaïbi, Monotonicity and stability of periodic polling models, *Queueing Systems*, **15** (1994) 211–238.
[13] S.W. Fuhrmann, A decomposition result for a class of polling models, *Queueing Systems* **11** (1992) 109–120.
[14] P.E. Gill, W. Murray and M.H. Wright, *Practical Optimization* (Academic Press, New York, 1981).
[15] W.P. Groenendijk, Waiting-time approximations for cyclic-service systems with mixed service strategies, in: M. Bonatti (Ed.), *Teletraffic Science for New Cost-Effective Systems, Networks and Services, ITC 12* (North-Holland, Amsterdam, 1989) 1434–1441.

[16] M. Hofri and K.W. Ross, On the optimal control of two queues with server set-up times and its analysis, *SIAM J. Comput.* **16** (1987) 399–419.
[17] G. Hooghiemstra, M. Keane and S. van de Ree, Power series for stationary distributions of coupled processor models, *SIAM J. Appl. Math.* **48** (1988) 1159–1166.
[18] J. Keilson and L.D. Servi, Oscillating random walk models for GI/G/1 vacation systems with Bernoulli schedules, *J. Appl. Probab.* **23** (1986) 790–802.
[19] P.J. Kuehn, Multi-queue systems with non-exhaustive cyclic service, *Bell Syst. Tech. J.* **58** (1979) 671–698.
[20] K.K. Leung, Cyclic-service systems with probabilistically-limited service, *IEEE Select. Areas Comm.* **9** (1991) 185–193.
[21] H. Levy and M. Sidi, Polling systems: applications, modeling and optimization, *IEEE Trans. Comm.* **38** (1990) 1750–1760.
[22] H. Levy, M. Sidi and O.J. Boxma, Dominance relations in polling systems, *Queueing Systems* **6** (1990) 155–171.
[23] Z. Liu, P. Nain and D. Towsley, On optimal polling policies, *Queueing Systems* **11** (1992) 59–83.
[24] I. Meilijson and U. Yechiali, On optimal right-of-way policies at a single-server station when insertion of idle times is permitted, *Stoch. Process. Appl.* **6** (1977) 23–52.
[25] R. Ramaswamy and L.D. Servi, The busy period of the M/G/1 vacation model with Bernoulli schedules, *Stoch. Models* **4** (1988) 507–521.
[26] S.S. Rao, *Optimization Theory and Applications* (Wiley Eastern Limited, 1978).
[27] J.A.C. Resing, Polling systems and multi-type branching processes, *Queueing Systems* **13** (1993) 409–426.
[28] L.D. Servi, Average delay approximation of M/G/1 cyclic service queues with Bernoulli schedules, *IEEE J. Select Areas Comm.* **4** (1986) 813–822.
[29] H. Takagi, Queueing analysis of polling models, in: H. Takagi (Ed.), *Stochastic Analysis of Computer Communication Systems* (North-Holland, Amsterdam, 1990) 267–318.
[30] T.E. Tedijanto, Nonexhaustive policies in polling systems and vacation models: qualitative and approximate approach, Ph.D. Thesis, University of Maryland, 1990.
[31] J.A. Weststrate and R.D. van der Mei, Waiting times in a two-queue model with exhaustive and Bernoulli service, *Z. Op. Res.*, to appear.
[32] U. Yechiali, Optimal dynamic control of polling systems, *Proc. 13th Int. Teletraffic Congress, Workshop: Queueing, Performance and Control in ATM* (North-Holland, Amsterdam, 1991) 195–208.

Hans Blanc was born in 1955 in Rotterdam. He received the Master's degree and the Ph.D. degree, both in Applied Mathematics, from the University of Utrecht, The Netherlands, in 1977 and in 1982, respectively. From 1982 to 1987 he was with the Centre for Mathematics and Computer Science in Amsterdam, Delft University of Technology and the University of Limburg in Maastricht, The Netherlands, successively. Since 1987 he has been with the Department of Econometrics, Tilburg University. His research interests include computer performance evaluation and optimization, queueing theory, and numerical analysis.



Robert D. van der Mei received his Master's degrees in Mathematics and Econometrics from the Free University of Amsterdam, The Netherlands, in 1990. Since December 1990 he has worked as a Ph.D. student at Tilburg University, The Netherlands. His research interests include queueing theory, numerical analysis, and performance evaluation.